GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

È difficile fare previsioni, specialmente sul futuro

… anche con il Machine Learning!

https://blog.minitab.com/blog/tough-making-predictions-about-future-machine-learning
Versione italiana a cura di Luca Biasibetti.

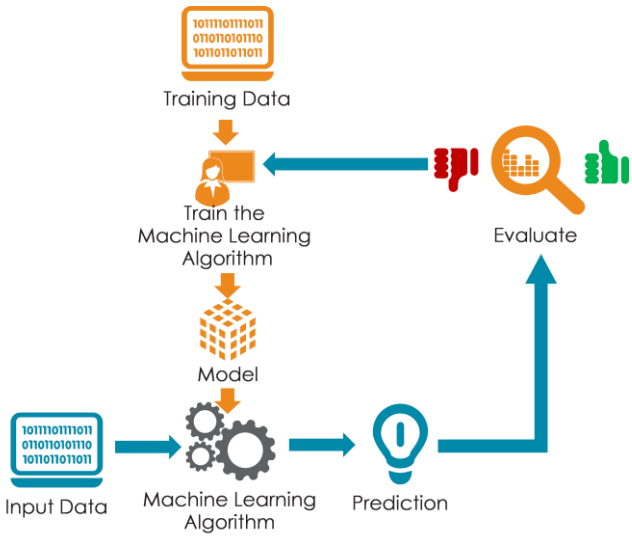
Bill Kahn guida lo statistical modeling group for consumer banking di Bank of America. Il suo team costruisce centinaia di modelli utilizzando un'ampia gamma di tecniche statistiche e machine learning. Questi modelli aiutano ad assicurare una stabilità finanziaria per individui, società e comunità distribuite su tutto il Paese. Negli ultimi decenni, Bill ha guidato gruppi di statistici in aziende di ambito finanziario, consulenza e manufatturiero. Ha conseguito una laurea in Fisica, un master in statistica a Berkeley e un dottorato in statistica a Yale.

Minitab ha chiesto a Bill di condividere le sue conoscenze sul Machine Learning (ML) come "base d'azione" nel mondo del business.

ALGORITMI MACHINE LEARNING

Tutti gli algoritmi di Machine Learning, in sostanza, seguono il seguente processo che può essere suddiviso in due parti. Nella prima parte, alcune sequenze di funzioni a complessità crescente si adattano a una parte dei dati (training data set). A questo punto, ogni modello nella sequenza è valutato in base alle performance che ottiene sui dati trattenuti in precedenza (holdout data set o set di controllo). Viene selezionato il modello con il più alto grado di adattabilità all'holdout data set. In questi passaggi ci sono diverse fonti di variabilità tra cui la sequenza esplorata, il modo in cui viene trovato ogni adattamento, la definizione di buon adattamento e quale holdout set viene selezionato in modo casuale. Si può concludere comunque che, con un paio di importanti

precauzioni, questa semplice sequenza produce in genere buone previsioni all'interno del campione.



DUE IMPORTANTI PRECAUZIONI

In primo luogo, dobbiamo utilizzare la giusta loss function per adattare e valutare i modelli. Se la loss function non è specificata in modo corretto tutti gli algoritmi di Machine Learning possono produrre risultati senza senso, come classificare tutti nel gruppo predominante e, di conseguenza, non riuscire a fare previsioni utili. È necessario utilizzare tutta l'esperienza possibile per selezionare una loss function che sia rilevante negli ambiti di interesse, ad esempio commerciale, scientifico e ingegneristico.
In secondo luogo, poiché ogni algoritmo possiede degli iperparametri (parametri che non possono essere stabiliti su basi puramente concettuali), dobbiamo esplorare un intervallo abbastanza ampio di questi per assicurarci che non stiamo utilizzando qualche "brutto" insieme di valori che potrebbe portare a decisioni deboli e inaccettabili.

PREDIRE IL FUTURO

Anche se il Machine Learning produce buone previsioni all'interno dei campioni analizzati, non è sempre ciò di cui abbiamo bisogno. Abbiamo bisogno anche di buone previsioni al di fuori del campione osservato. Questo salto, dallo studio dell'esperienza passata a quello del comportamento futuro, è molto grande e richiede alcune considerazioni aggiuntive guidate da fondamentali principi statistici. Queste considerazioni includono: la selezione del giusto problema, la selezione di variabili dipendenti significative, la segnalazione del bias dei dati sottostante, la comprensione della gerarchia e della dipendenza tra le osservazioni e la costruzione della giusta sequenza di modelli. Nessuno tra questi requisiti è esclusivo per il Machine Learning: tutti sono essenziali affinché qualsiasi analisi statistica sia affidabile.

Ciò che di meglio può fare un modello, è estrarre le informazioni insite nei dati. Affinchè i dati contengano informazioni preziose, il design sperimentale nei modelli Machine Learning è molto importante così come è importante per qualsiasi altro modello predittivo. Dopo aver eseguito un progetto ben strutturato, è possibile creare un modello Machine Learning e assegnare un "punteggio" a ogni osservazione per ogni possibile combinazione di variabili controllabili. Quello che bisogna imparare è come impostare gli input controllabili (ad esempio, prezzo, canale di marketing, temperatura o velocità) per produrre il miglior risultato possibile per ognuno di essi.
Questo approccio sfrutta al massimo ciò che conosciamo, ma presenta uno svantaggio significativo per qualsiasi sistema il cui stato reale evolva nel tempo. Il miglior settaggio dei fattori di controllo varia quando l'ambiente esogeno finisce alla deriva (ad esempio, l'evoluzione della qualità delle materie prime, la contropartita dei consumatori o la risposta della concorrenza). Se dovessimo sempre fare una singola assegnazione ottimale per un particolare input, confonderemmo ciò che osserviamo con ciò che facciamo. Questo renderebbe impossibile la creazione di un modello nuovo e migliorato. I nuovi disegni sperimentali, come il campionamento di Thompson, risolvono il problema "sfidando" continuamente le convinzioni acquisite. Questi progetti consentono di raggiungere un equilibrio ottimale tra l'ottenimento di un guadagno nel presente e l'apprendimento di ciò di cui abbiamo bisogno per essere in grado di guadagnare anche in futuro.

Il Machine Learning è una potente aggiunta ai normali strumenti posseduti dai professionisti del settore. Con una certa cautela di base, che consente di evitare ciò che è sciocco o terribile, in combinazione con una serie di abilità statistiche classiche, il Machine Learning aiuta a rendere gli statistici sempre più efficienti e performanti.